All our collective ingenuity will be needed

PAUL BERG

Department of Biochemistry, Stanford University Medical Center, Stanford, California 94305, USA

FROM ITS INCEPTION AND SUBSEQUENT DISCLOSURE to the scientific public, the proposal to sequence the human genome has had staunch advocates and vehement detractors. As conceived, the project's goal was to obtain the sequence of the 3 billion base pairs comprising the human haploid genetic complement. But the debate among scientists and a realistic cost benefit analysis revised the goals of the project in several significant ways. One was to include a moderate resolution map of linked RFLP markers for use in locating disease genes; a second was to obtain a physical map of cloned DNA segments spanning the entire genome. Moreover, it soon became apparent that the sequencing efforts would have to await completion of the first two objectives as well as the development of speedier, more accurate, and far less expensive sequencing technologies. Such a reformulation of the project was inevitable once fine scientific minds abandoned their prejudices and turned to developing a coherent and workable strategy.

The revised timetable and early emphasis on obtaining genetic and physical maps will have its most immediate effect on locating and identifying genes responsible for both single and multigenic disorders. Undoubtedly the pace at which genes associated with a variety of diseases will be mapped and isolated will quicken substantially, thereby hastening the day when their structure, function, and disease relevance can be analyzed in molecular terms.

Another major modification of the original plan was to increase the number of genomes to be included in the project. Rather than 'pork barreling,' this expansion of the project recognizes the close correspondence in genetic structures and functions between organisms even distantly related. Furthermore, it is evident that such relatedness will inform and speed the work on the human genome. In addition, the research with yeast, *Drosophila*, nematode, and mouse genomes provides experimental models with which hypotheses and technologies can be tested without resorting to human experimentation.

Despite the revisions in the program's goals and timetable, the debate about the project's legitimacy and value lingers and threatens to polarize the constituencies needed for the project's support. One of the principal reservations concerns the cost and style of the project, particularly the perception that both challenge the traditional mechanisms for supporting investigator-initiated research in broad areas of biology, and even in areas related to but excluded from the genome project. The simplistic view of this challenge is that the genome project represents "Big Science vs. Little Science." But aside from the formulation of specific goals by the Genome Project's Advisory Committee, i.e., creating genetic and physical maps, improving instrumentation and informatics capability, and obtaining genomic nucleotide sequences, the means for achieving those goals remains via research proposals initiated by single or relatively small groups of investigators. The more appropriate concern, it seems to me, is whether the direction of the project will be heavy- or light-handed, i.e., whether the evaluation of research proposals will be too narrowly construed with respect to the ultimate aims of the project. In the long run, however, the early genome mapping studies, the availability of clone banks, the development of rapid and cheaper sequencing and nucleic acid synthesis technologies, and vastly improved informatics will greatly benefit the research activities of all biological scientists.

". . . that 95% of the mammalian genome's sequence is junk expresses a prejudiced definition of genes."

The second and perhaps more disturbing threat to the project's viability stems from the still-divided views concerning its scientific merit, particularly the value of the sequence information relative to the cost of obtaining it. An often-cited view is that the only genetically relevant sequences in mammalian genomes is the 5% that encodes protein chains, and in a few cases, mature RNA species. By inference, often stated as fact, the remaining 95% is junk and not worth knowing!

Two points are worth making in this context. First, if our goal is to understand genome function in molecular terms, then it seems myopic to regard the protein and RNA coding sequences as the only relevant information. Are there any doubts that DNA replication, chromosomal packaging and segregation, and a variety of other DNA transactions essential to life are specified and guided by the genome's sequence? Should they be regarded as junk at this stage of our ignorance? Second, perpetuation of the supposition that 95% of the mammalian genome's sequence is junk expresses a prejudiced definition of genes. A more appropriate view of a gene is that it consists of the entire transcription unit (exons plus introns) and the regulatory regions involved in governing that unit's expression under a variety of conditions. By that definition, current data suggest that about 50% of the genome's sequence is comprised of genes. There is already clear evidence that specific sequences in introns and in intergenic regions constitute important regulatory signals. Shall we foreclose on the likelihood that the so-called noncoding regions within and surrounding genes contain signals that we have not yet recognized or learned to assay? Are we prepared to dismiss the likelihood of surprises that could emerge from viewing sequence arrangements over megabase rather than kilobase distances?

Achievement of the Genome Project's goals may mark the end of the program, but it presages the beginning of a new era of study and understanding of humans as a biological entity. All of our collective ingenuity will be needed to learn the mechanisms by which the genome's sequences specify our species' characteristic phenotype.

0892-6638/91/0005-0075/$01.50.